Introduction: Einführung

Paarsch, Harry J ; Schmedders, Karl

Abstract: By making gathering large samples of data (Big Data) almost trivial, the Information Revolution has changed fundamentally how many scientists now conduct empirical research. The explosion in the variety and volume of information that is Big Data has in many cases altered both the questions asked and how those questions are answered. In this special issue devoted to Big Data, we have collected five papers from the social sciences, particularly economics, but business as well. The main goal of the issue is to introduce economists to the different ways that Big Data can and have been used in business and economic research, in the hope that this will spur additional innovative research in those fields.

DOI: https://doi.org/10.1515/jbnst-2018-0028

Editorial

Harry J. Paarsch and Karl Schmedders*

Introduction

Einführung

https://doi.org/10.1515/jbnst-2018-0028

Abstract: By making gathering large samples of data (Big Data) almost trivial, the Information Revolution has changed fundamentally how many scientists now conduct empirical research. The explosion in the variety and volume of information that is Big Data has in many cases altered both the questions asked and how those questions are answered. In this special issue devoted to Big Data, we have collected five papers from the social sciences, particularly economics, but business as well. The main goal of the issue is to introduce economists to the different ways that Big Data can and have been used in business and economic research, in the hope that this will spur additional innovative research in those fields.

Keywords: big data, data science
JEL Classification: C1, C8, C9

1 Motivation and goals

By making gathering large samples of data (Big Data) almost trivial, the Information Revolution has changed fundamentally how many scientists now conduct empirical research. The explosion in the variety and volume of information that is Big Data has in many cases altered both the questions asked and how those questions are answered. Although Big Data have natural applications in biology and medicine (for example, when examining human genes), in this special issue we have collected five papers from the social sciences, particularly economics, but business as well. The contributions cover a variety of different topics—including A/B testing and copula estimation, too—in such diverse fields as heavy-equipment rentals as well as banking and finance. The contributors are

***Corresponding author: Karl Schmedders,** Universitat Zurich Institut fur Betriebswirtschaftslehre, Moussonstrasse 15, Zurich, Switzerland, E-mail: Karl.Schmedders@business.UZH.ch
Harry J. Paarsch, University of Central Florida College of Business Administration, Orlando, USA, E-mail: Harry.Paarsch@UCF.edu

equally diverse—being employed in private industry at such firms as GoDaddy in Kirkland, Washington and Microsoft in Seattle; government at such agencies as Payments Canada and Statistics Canada, in Ottawa; and industry groups such as the American Rental Association in Washington, DC; and six different academic institutions—Carleton, Emory, Georgetown, Lakehead, and Stanford universities as well as the University of California at Santa Cruz.

The main goal of this special issue is to introduce economists to the different ways that Big Data can and have been used in business and economic research, in the hope that this will spur additional innovative research in those fields.

2 The papers and their authors

This issue contains five papers: The first is a descriptive paper by Jay Dixon (Statistics Canada), Robert Petrunia (Lakehead University), and Anne-Marie Rollin (Statistics Canada) that demonstrates the virtues of large administrative databases using an example replete with firm-specific heterogeneity gleaned from the databases at Statistics Canada. In the second paper, Konstantin Golyaev (Microsoft) discusses perhaps the most interesting feature of data collection on the Internet: randomized experiments. Jason Ansel (GoDaddy), Han Hong (Stanford University), and Jessie Li (University of California at Santa Cruz), in the third paper, provide some clever and interesting insights concerning how to use data from randomized and conditionlly randomized experiments effectively, and then apply those insights to data gathered from the GoDaddy auction site on which Internet domains are sold. In the fourth paper, John McClelland (American Rental Association) and John Rust (Georgetown University) illustrate how Big Data can be levered to improve the distribution of heavy-equipment rentals across firms, in an effort to improve profits; that is, they implement stochastic dynamic programming at scale. In the fifth and final paper, Leonard Sabetti (Payments Canada), David T. Jacho-Chávez (Emory University), Petrunia, and Marcel Voia-Cristian (Carleton University) demonstrate how Big Data can be used to assess credit risk in the automated clearing settlement system of Canada—a system that produces a tsunami of data each business day.

In short, we believe that the papers collected below represent a fine blend of both theory and application.

2.1 Large administrative databases

Large administrative databases represent a wealth of information concerning both businesses and citizens. Dixon, Petrunia, and Rollins illustrate how to use

business tax data from Statistics Canada's Longitudinal Employment Analysis Program database to describe the annual firm growth rate distribution in Canada during the 2000–2009 period. Administrative tax databases are essential when investigating firm growth because the information contained in them permit researchers to examine the universe of all employer firms and thus to capture the entire firm growth distribution. Their results suggest that the distributions of employment growth rates in Canada have more mass both in the center and in the tails than the Gaussian law. This evidence paints a picture of firm growth dynamics whereby a large number of firms change very little each year, and a small number markedly decline or grow—results that hold across industries and time as well as age and size of firms.

2.2 Randomized experiments

Perhaps the greatest contribution of statisticians to science in the twentieth century has been randomized controlled experiments, sometimes referred to in business as A/B testing. Although the American logician and philosopher Charles S. Peirce, along with the Polish psychologist Joseph Jastrow, introduced and implemented blind, randomized, controlled experiments in Peirce and Jastrow (1885), it was the English biologist and statistician Sir Ronald A. Fisher who did the most of any scholar to further this research agenda, especially through his influential book *The Design of Experiments* Fisher (1935), which provided a guide to implementing the method throughout applied science.

Recently, with the advent of Big Data, many ingenius applications of A/B testing have proven valuable in business, particularly ecommerce. Implementing randomization on a computer is a tricky business, as Golyaev points out. Specifically, practical considerations related to business as well as engineering realities related to latencies result in compromises. Such compromises sometimes lead to deviations from Fisher's ideal framework. Golyaev demonstrates the potential effects that such compromises can have.

2.3 Some theory and an application

Ansel, Hong, and Li investigate estimation and inference of the average treatment effect parameter when a binary instrumental variable is generated by a randomized experiment. They first show that adding covariates and their interactions with the instrument improves estimation precision of the local average treatment effect. Unfortunately, the conventional robust two-stage least-squares (2SLS) standard errors are no longer valid—either an analytic correction or bootstrap

resampling need to be employed. Ansel, Hong, and Li also examine inference concerning a finite sample parameter, such as the average treatment effect for the in-sample compliers. In this case, the robust 2SLS standard errors are typically only conservatively valid, but covariates can be used to reduce the length of the condence intervals. Next, they generalize their results to different randomization schemes including stratified randomization and covariate adaptive randomization. Finally, the authors demonstrate how all of these methods can be used to investigate data from actual field experiments conducted by the firm GoDaddy where Internet domain names are sold at auction.

2.4 Heavy-equipment rentals

In economics as well as operations research, the problem of when to replace an asset has been studied since at least the nineteenth century. In forestry, the work of the German forester Martin Faustmann is an obvious case in point, but examples from other industries, particularly those involving machinery, have been the focus of much research since the 1950s. Typically, machine replacement policies are analyzed as solutions to dynamic optimization problems because the optimization occurs over time, and decision making is sequential: successive decisions are conditioned on information that changes over time—such as the age of the machine, its physical condition and resale price, and the state of the business cycle. Dynamic optimization problems invariably account for randomness in rental revenues as well as maintenance and replacement costs, and resale or scrap values. Although characterizing a solution to such problems using the method of dynamic programming has been well-known for at least a half a century, a dearth of relevant data and slow computing hardware has made implementing such policies impossible, except in the case of small examples. Using Big Data and harnessing modern computers, McClelland and Rust demonstrate that what was initially a theoretical curiosity is now a practical reality—namely, solving for the optimal policy function for potentially thousands of different machines in a flexible, personalized way.

2.5 Assessing credit risk

In the final paper, Sabetti, Jacho-Chávez, Petrunia, and Voia study a credit risk (collateral) management scheme for the Canadian retail payment system designed to cover the exposure of a defaulting member. They estimate the *ex ante* size of a collateral pool large enough to cover exposure for a historical worst-case

default scenario—employing two statistical approaches based on extreme value models, which require large amounts of data to estimate extreme order-statistics reliably. Their empirical results (that the largest daily net debit position across participants exceeds $1.5 billion once a year) are informative for optimal collateral management and system design of pre-funded retail-payment schemes.

Acknowledgment: We thank Peter Winker, Managing Editor of the *Journal of Economics and Statistics* / *Jahrbücher für Nationalökonomie und Statistik* for inviting us to be guest editors of this Special Issue on "Big Data" as well as the contributing authors for making the task so pleasant.

References

Peirce, Charles S. and Joseph Jastrow (1885), On Small Differences in Sensation. Memoirs of the National Academy of Sciences 3: 73–83.

Fisher, Ronald A. (1935), The Design of Experiments. Edinburgh, UK, Oliver and Boyd.

Article note: This article is part of the special issue "Big Data" published in the Journal of Economics and Statistics. Access to further articles of this special issue can be obtained at www.degruyter.com/journals/jbnst.